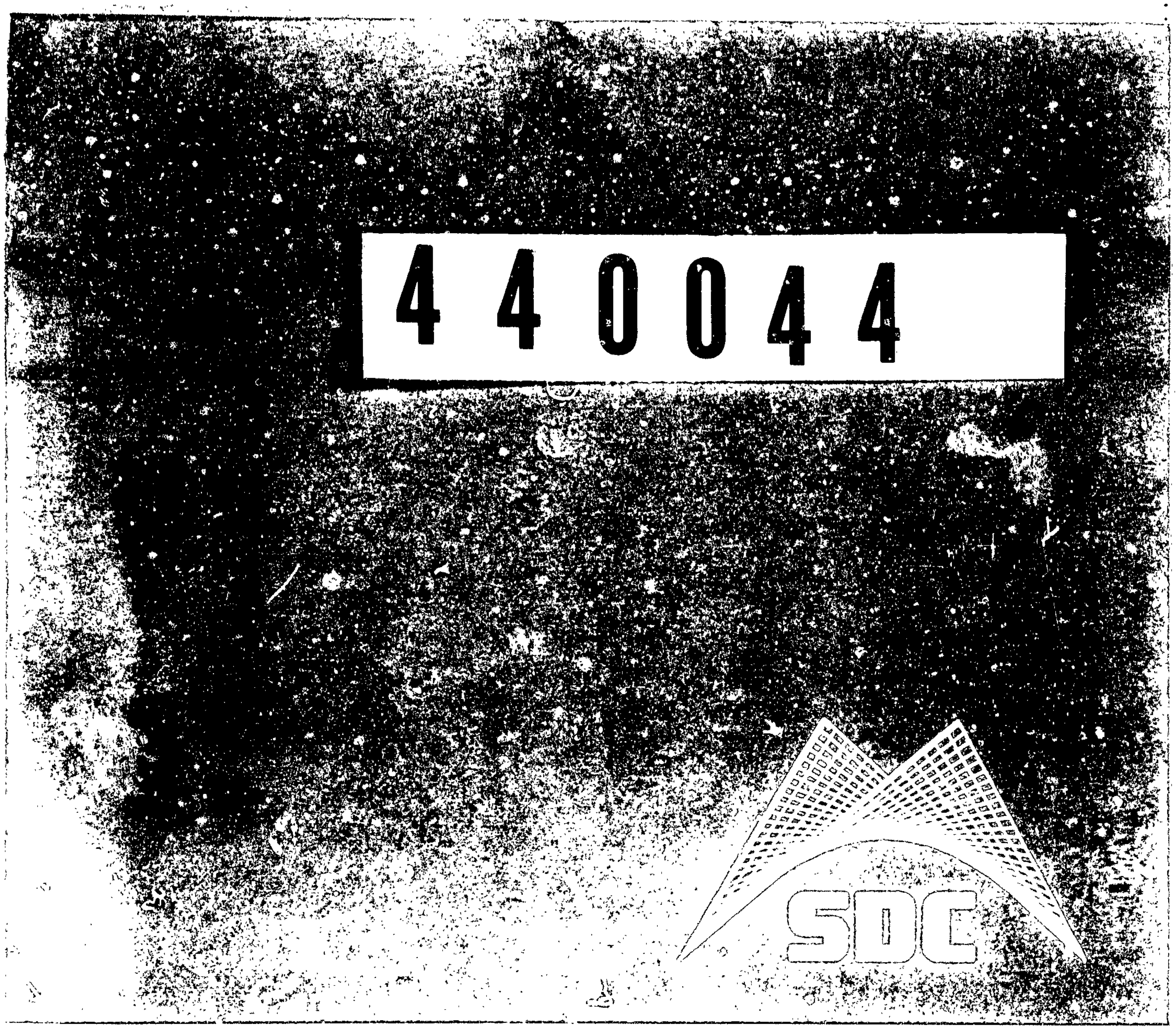
AD NUMBER
AD440044
NEW LIMITATION CHANGE
TO Approved for public release, distribution unlimited
FROM Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; MAR 1964. Other requests shall be referred to US Air Force Personnel Research Laboratory, Attn: Skill Components Research Laboratory, Lackland AFB, TX.
AUTHORITY
SDC ltr, 31 Jul 1969

NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.



SP-1481

Some Compromises Between Word Grouping and Document Grouping

Lauren B. Doyle

26 March 1964

(SP Series)

SP-1481

Some Compromises Between Word Grouping and Document Grouping

by

Lauren B. Doyle

26 March 1964

SYSTEM DEVELOPMENT CORPORATION, SANTA MONICA, CALIFORNIA

This paper was presented at the Symposium on Statistical Association Methods for Mechanized Documentation held March 17 through 19, 1964, at the National Bureau of Standards, Washington, D. C.

ABSTRACT

Statistical analysis of the text of document collections has yielded, for information retrieval purposes, two broad classes of output: word grouping and document grouping. Associative indexing comes under the general heading of word grouping; automatic classification is a kind of document grouping. Document grouping and word grouping, however, can be combined to give a scheme of classification with more attractive features than could be achieved with either document grouping or word grouping alone.

A hierarchical grouping program written by Joe H. Ward of Lackland Air Force Base for use in classifying personnel by skill and aptitude turns out to be nearly ideal as a basis for a mixed document and word-grouping approach. The program will derive four-or five-level hierarchies from keyword lists drawn from 100 documents, will position document numbers or other numbers in the smallest subcategories, and is capable, with additional routines, of extracting appropriate labels from the keyword lists to describe the categories at all levels of the hierarchy. Additionally, homograph separation occurs as a natural outcome of the program's operation.

SOME COMPROMISES BETWEEN WORD GROUPING AND DOCUMENT GROUPING

Information retrieval technology in the 1950's was based largely on principles of logic,* an emphasis that was perhaps a "logical" result of the emphasis on use of computers in information retrieval. Computers are (above all) logical. Then a well-known logician (1) said that logic was at least being grossly misapplied or was at worst nearly useless in the information retrieval field.

Judging by the interest in statistical approaches in general and associative indexing in particular, the 1960's will see information retrieval based more and more on principles of redundancy. This is more appropriate because, as we are often so painfully aware, the literature is quite redundant and not very logical.

Redundancy has the adverse connotations of undue length and repetition. It is these very characteristics that make a statistical approach to text analysis and retrieval both feasible and desirable. Undue length favors a statistical approach becuase it increases the sample size and, needless to say, the world's technical literature is unduly sizable as a sample. Repetition, of course, gives us something to count, without which we would have no statistics; but more important than that, selective repetition by authors can be a highly reliable clue to topic, as recognized by H. P. Luhn (2).

Document Grouping

Those who try to employ redundancy as a means of automatically generating an organized structure giving us access to the literature do so in two general ways--by document grouping and by word grouping. Document grouping was the basis of library classification long before there were computers, and it is expectable that the statistically orientated would try to duplicate by automatic means what the librarian can do intellectually, because similarity of word content in a group of documents implies similarity of topic. Of course, documents or references thereto (titles, etc.) can be grouped in other ways than by word content similarity; as examples, permuted-title indexing groups them alphabetically, and citation indexing groups according to author-implanted cues. These approaches currently outrun the statistical approach in popularity because, among other things, they are cheaper; neither method requires the entire text of an article to be processed or additional intellectual work to be done other than that done by the author himself.

*Mainly the principles of Boolean algebra.

But we value the statistical approach, in spite of its current expense, not only because costs are rapidly declining and will result inevitably in feasible digital storage for entire documents, but also because it is a whole technology, whose applications to text analysis go beyond what we talk about here. As one example of that, statistics can be shown to be a strong right arm for syntactic analysis (3), and perhaps--eventually--for machine translation. This is so because the redundancy in text manifests itself through the grouping of words as well as through the grouping of documents.

Word Grouping

In my own work I have been preoccupied with word grouping (4). Others, such as H. E. Stiles (5), have in effect used index-term grouping, which is equivalent to word grouping, to improve the performance of literature-searching systems. Words or terms can be grouped statistically as a result of their hig co-occurrence in the same documents as tags or keywords; when co-occurrence is high, as measured by some statistic, we speak of the co-occurring words as being strongly "associated." Both word grouping and document grouping spring from the tendencies of many words to co-occur strongly.

Developments in statistical word association are proceeding along two paths. The majority approach is that of Stiles, which is a modified coordination indexing in which users formulate search requests and in which the machine acts on those requests in such a way that the retrieved documents contain not only the words specified by the request, but also words that are associated statistically to those in the request.

The second approach, which is still a rather small minority, is that in which the computer is used to generate an "association map" as a printout or cathode-ray tube display. The best way to visualize the difference between these two approaches is by analogy to the difference between straight machine searching of text and automatic indexing. In machine searching, one makes a request that is fed into the machine as a criterion the machine can use in searching for relevant references. In automatic indexing, the machine is used not as a searching instrument but as an arranger of references that can be scanned in printout form by the human eye. In associative indexing, by analogy, the first approach involves user specification of what the machine should look for and the second approach generates a printout or display by which the user himself can search.

The analogy here might even extend to developmental history. Recent years have seen a shift away from machine searching toward automatic indexing, especially permuted-title indexing. We might well be on the point of shifting from machine-associational searching to machine-associative indexing. I am assuming so, and have for this reason habitually placed my eggs in that basket.

Association maps can take on a bewildering variety of forms. The forms with which I have become most familiar are shown in Figure 1; on the left is a map hand-drawn from computer-generated statistical co-occurrence data, and on the right is a "hierarchical map" generated from the same text as the map at left. Both of these forms were first discussed in 1961 (6) and both are capable of completely automatic generation from text. The map at left could be called a "raw-association map," in that it faithfully reflects the most strongly co-occurring word pairs in the corpus; the hierarchical map at right sacrifices strong co-occurrences between words of roughly equal frequency for the sake of better organization. The hierarchical effect is achieved by discriminating against relating (linking) words of more or less equal frequency and by relating words of high frequency* to words of lower frequency in a cascade of categories and subcategories, as shown; since the words of high frequency apply to a larger number of documents, it follows that these would be used to label the larger categories. One can construct ad hoc statistical functions by which one can bring about the desired discrimination against co-occurrences between equally frequent words. The most effective one I have found so far is:

$$F = \frac{\frac{2c}{b} - 1}{(b/a - 0.35)^2 + 0.03},$$

where a = the value of the higher frequency, b = the value of the lower frequency and c = the frequency of co-occurrence of words a and b. The numerator's purpose is to maximize F as documents with tokens of word b as tags or keywords approach 100% inclusion in the larger set of documents having word a. The denominator maximizes F as the ratio of the two frequencies, a and b, approaches 0.35; such a function would thereupon favor hierarchies having on the average three subcategories per category The presence of the constant 0.03 in the denominator is to prevent the function from approaching infinity.

Disadvantages of Pure Word or Document Grouping

The reason I now search for compromises between word grouping and document grouping is that I have become aware of certain disadvantages to using either approach in a pure way. Pure document grouping, for example, suffers from two weaknesses:

*By "frequency" here we mean "number of documents having this word or tag," rather than "number of words." The author, in a previous article (4), has defined this kind of frequency as "prevalence."

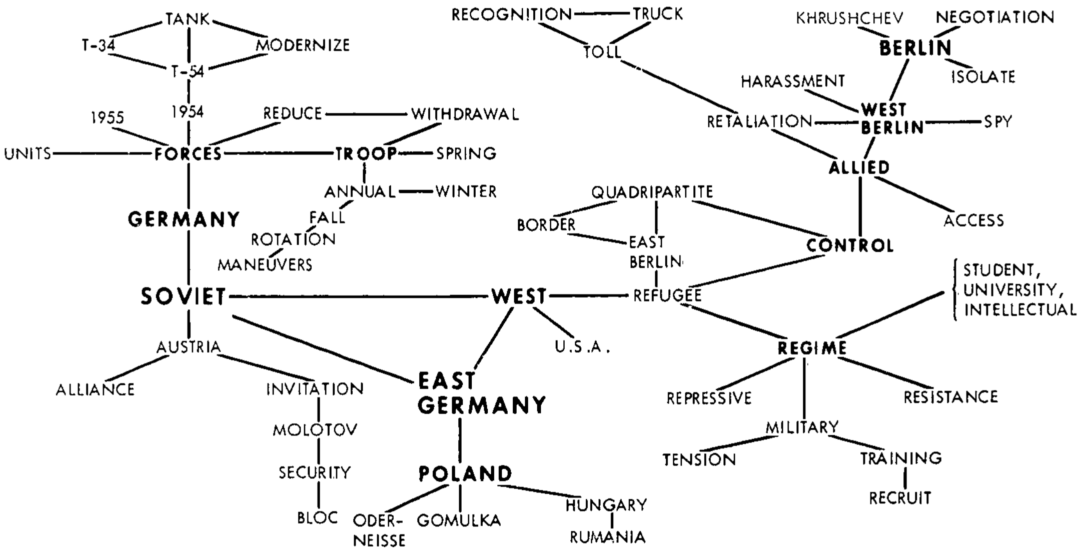


Figure 1A. Association Map.

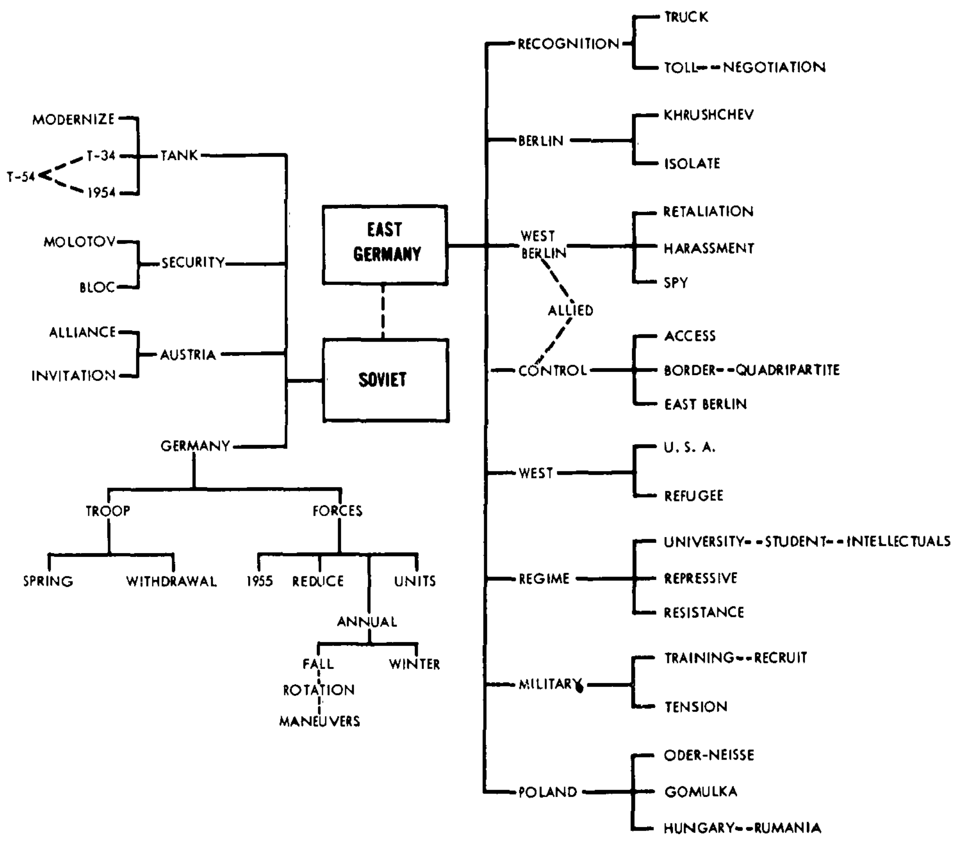


Figure 1B. Hierarchical Association Map.

1) There is no obvious clear-cut way to represent the groups of documents for perusal by literature searchers. Grouping of titles in correspondence to the document groups is not entirely adequate, because the similarities leading to group formation may not be evident, and because a flock of titles may contain too much information to characterize whole groups, leading to cognitive strain for searchers who would like to inspect numerous groups.

2) The organization of the groups themselves, though potentially achievable automatically, may not be representable in a scheme that can be followed by a searcher.

These faults would not seem important to those who take the viewpoint of Maron (7) and others, which pictures "heuristics in document space" as a means of machine retrieval of closely related documents. These workers would not be inclined to emphasize representation for search by the human eye.

Word grouping (association maps, hierarchical maps) has three weaknesses as a pure approach:

1) Since the basic idea of an index based on word groups is to find word clusters of interest or pertinence and to proceed from such a cluster to references containing more information about the documents whose co-occurring words caused the cluster, it is important that word maps have document numbers (or other indicators) positioned properly on them. This proves difficult to do reliably by automatic means.

2) Homographs are a problem in word-grouping techniques. Though statistical separation of homographs has been shown to be feasible by Stiles (8), it ordinarily would require that an additional statistical technique be used along with whatever is used for the word grouping. We would like to find a statistical technique from which both word grouping and homograph separation come in natural consequence.

3) Though word grouping (particularly the "hierarchical map") suggests organization of something, the literature searcher is given no sense of what it is that has been organized. A map, in order for one to accept it as a meaningful entity, ought to be a map "of something." An organized set of document clusters, if it can be represented in a map-like way, would have much more reality to a searcher because it would be perceived as a map of the document collection.

A Procedure Permitting Both Document and Word Grouping

I could not have expected that these grim doubts about either document grouping or word grouping could be cleared up by a single computer program used in a field quite remote from document retrieval. However, early in 1963 an article by Ward and Hook (9) came to my attention that described a hierarchical grouping procedure used by the United States Air Force in grouping aptitude profiles for personnel assignment. I was fortunate enough to obtain the corresponding Fortran II computer program, which was implemented and run on our Philco 2000. I used this program, in effect, as a document grouping program.

As a natural outgrowth, perhaps, of my preferred orientation toward word grouping, I found that one can superimpose a highly organized word pattern on the document grouping pattern that the program generates, and that this superimposed word pattern not only describes the document groups, but also overcomes the three weaknesses of a "pure word-grouping" approach.

I do not have the time or inclination to discuss the mathematical principles of the grouping program, which are described well enough in the Ward paper (9). Adherents of this statistical approach spend much time arguing among themselves as to whether this or that statistical technique is more appropriate, but those who have a chance to compare them (10) often find that the difference in output between one technique and another is not appreciable. Indeed, even if one technique led to a substantially different output than another, it would be hard to say that one result was right and the other wrong. I have usually found that selection of technique on purely mathematical grounds is appropriate only when there is full and complete understanding of what the technique is supposed to do; otherwise the only sensible thing to do is to base selection of technique on an appraisal of the utility and quality of output. This can lead to the choice of a completely ad hoc statistic with no foundatior in mathematical theory, as in the case of the hierarchical map shown in Figure 1.

Several runs of the Ward program were made, each having 100 twelve-word lists as input. Each twelve-word list can be regarded as a list of index tags or most-frequent content words of one document. The output, then, can be viewed as the organization by similarity of a 100-document library. Three runs will be described herein, one on 100 lists corresponding to reports on German affairs, one on 100 lists corresponding to information retrieval papers, and 100 that include 50 lists each from German affairs and physics collections.

Principle of Operation of Ward's Grouping Procedure

Before presenting the results of these computer runs, it is desirable to give a non-mathematical description of how the program operates. Its objective is to form groups whose members have maximal similarity to each other. In the runs described above, it begins with 100 ungrouped lists or, it would be better to say, 100 groups having one member each. Each program "pass" forms one group of two members or more according to any of the following three rules:

1) Combine one list and another list to form a group of two lists.

2) Add one list to a group of two or more lists.

3) Merge two groups of two or more lists.

Note that never more than two entities (lists or groups of lists) are combined on a given pass; therefore any one pass diminishes the total number of groups (remembering that we've designated ungrouped lists as "groups with one member only") by one; and also, therefore, the total number of passes must be n-1 for a collection of n lists. In other words, the program accepts n lists as input, forms a new group (in accordance with the rules just given) in each of n-1 passes, and on the (n-1)th pass forms one large group consisting of all n lists.

There are, of course, a large number of paths that the program could follow to reach the all-inclusive group at the (n-1)th pass. For example, for a collection of four lists two possible paths exist if we think of the lists as indistinguishable: 1) form two groups of two each and merge these to form a group of four; or 2) form a group of two, add a third, and add a fourth. When we introduce combinations, however (i.e., regard the lists as distinguishable and count all possible ways of combining them), we find that the program has 18 possible paths by which to achieve the final group of four. On the first pass it can form any of six possible groups of two. On the second pass it can--for each of the six possible pairs--do three things: 1) group the two ungrouped lists, 2) add one of the ungrouped lists to form a group of three, or 3) add the other of the ungrouped lists to form a group of three. On the third pass all roads lead to Rome, i.e., the final group of four.

As the number of items to be grouped increases, the number of possible paths the program is allowed to take increases enormously. According to an earlier report of Ward's (11), for a group of five there are 180 possible paths, for six, 2700, for seven, 56,700, and for eight, 1,587,600.

The essence of Ward's grouping procedure is that out of the $\frac{(n!)^2}{n(2^{n-1})}$ possible paths for n items, it selects some one pathway that will most quickly bring together the items of greatest similarity. This selection is not as difficult as it may seem. Each of the (n-1) iterations is involved in selecting the total pathway, for on each program pass a group is formed such that the following function is maximized:

$$F = A_0(n_0-1) - A_1(n_1-1) - A_2(n_2-1) - C$$

In this function, n_o stands for the size of the group that is a candidate for formation on a given pass. On the first pass n_o must equal 2. On later passes the upper limit of n_o is the number of the pass plus one; the lower limit, however, is always 2 except on the final pass, wher n_o must equal n. The n_1 and n_2 are the sizes of the groups to be merged on a given pass, and their values are restricted by the relation $n_o = n_1 + n_2$, with a lower limit of +1 for either or both.

A_o , A_1, and A_2 are the corresponding average similarities for the groups, which we define as $A = \frac{\Sigma x}{n(n-1)/2}$, n in this case being the group size and x being some measure of the similarity of two of the items (in the case of the word lists used in this study, x was simply the number of words that two lists have in common). The summation of x includes all combinations of the n items taken two at a time. C is an arbitrary constant usually set at the maximum possible A value.

The above function, F, acts in effect as a threshold, being set at its highest achievable value at the beginning of the first pass. Highest achievable value means here that only items identical in all respects could be formed into groups. If all n items of a collection were identical to each other, the threshold F need never be lowered. But in that case, of course, there would be no point in forming groups.

In a typical collection of complex items, no two of which are identical, the program lowers the value of threshold F until two items are found similar enough to each other to constitute the "most similar pair in the collection." After the first pair is formed, the role of F becomes more complicated--and correspondingly more difficult to describe. For a comprehensive mathematical explanation, one should consult the Ward article (9). I have described the function to this extent only for the benefit of those who might want to construct their own grouping algorithm without having to decipher what in some cases might prove to be unfamiliar mathematical notation.

It will suffice for the purposes of this paper to state that F's role is to select at any given pass that group which has the most satisfying blend of similarity and homogeneity. The Ward program contains an alternative mode in which groups are formed solely on maximum average similarity; however my experience with this mode has convinced me that better classification is achieved (for my material, at least) in the mode that maximizes F, rather than average similarity of the next-to-be-formed group (i.e., A_o). Close scrutiny of F will show the reader that a candidate for group formation is penalized to the extent that the average similarity of the new group differs from the average similarities of the component groups. This has the result that on many passes groups are formed whose A_o values are substantially less than the maximum possible on those passes.

Intuitive Explanation of the Ward Procedure

The practical significance of the grouping procedure described can be better understood if we think about the problems involved in grouping common objects in terms of their attributes. Suppose, for example, that we apply the three rules given at the beginning of this section to forming groups from four objects: a plum, a walnut, a flowerpot, and a jar of mustard. Without splitting too many hairs on the question of specifying their attributes, it might seems reasonable to group the walnut and the plum first because they are both small, edible, tree-grown objects; furthermore, even without a knowledge of biology we suspect that they have many more things in common than we could perceive with the eye.

The next question is what to do on the second pass. There are two things that can be done. One (grouping the flowerpot with the plum and the walnut) appears unreasonable, since the flowerpot has practically nothing in common with either of the other two. The jar of mustard, however, can either be grouped with the flowerpot (because it is a non-metallic container, which just happens to contain mustard), or it can be grouped with the walnut and the plum (because it has in common with them the quality of being partly edible--the edible part being likewise derived primarily from vegetable sources).

Which of the above choices we would make depends on which attributes are of greatest interest to us. For example, if we were running a store we would unquestionably want to group the edibles, whereas if we were in the transportation business we would tend to group jars of mustard with flower pots because they present fewer problems in handling than do the perishable walnuts and plums.

Coming now to the world of document retrieval, how would we want to group books about walnuts, plums, flowerpots, and jars of mustard? Of course, a lot depends here on the aspects of these four subjects that are being discussed--for example, plums can be discussed as crops or as plants (under biology or botany). It is to be noted, however, that since jars of mustard and flowerpots are finished products, it is somewhat more difficult to think of any book that might treat them in a scientific (i.e., natural science) light, whereas any book "all about walnuts" or "all about plums" would of necessity have to begin with a biological discussion. From a librarian's viewpoint, then, it might be logical to group a book "all about jars of mustard" with similar books under the topic "manufacturing." A book all about flowerpots would probably also be found under the "manufacturing" heading, though not specifically in the area of food processing.

Fortunately, in the area of statistical methods of classification, we don't (yet) have to worry about such hard intellectual choices as the above librarian might have to make; at this point, we have nothing better than the simple and somewhat comfortable hypothesis that documents containing similar quantities of roughly the same words must be on roughly the same topic. This makes it quite easy for us to decide how we want things to be grouped.

In particular, it was easy for me to decide by what criteria I wished to group the twelve-word lists (described above)--group lists according to the number of words held in common. Let us assume that, based on a word count of books about walnuts, plums, flowerpots, and jars of mustard, I have derived* the following twelve-word lists:

1) WALNUT
TREE
NUT
HULL
SPECIES
WOOD
SHELL
LUMBER
KERNEL
BLACK
CROPS
SOIL

2) PLUM
FRUIT
SPECIES
TREE
PLANT
COLOR
GROW
BLOSSOM
SOIL
PRUNE
PIT
HYBRID

3) CLAY
PLANT
POT
MOLD
FIRE
POTTERY
DRY
COLOR
HEAT
HORTICULTURE
HOME
FLOWER

4) MUSTARD
SEED
BOTTLING
BLEND
SPICE
VINEGAR
PROCESS
FLOUR
SPREAD
FLAVOR
SANDWICH
SHARPNESS

One now notes that lists 1) and 2) have three words in common ("tree," "soil," and "species"), and that lists 2) and 3) have two words in common ("plant" and "color"). List 4) has no words in common with any of the others.

The outcome of our grouping procedure would be that the first program pass would group lists 1) and 2). The second pass has no choice but to put list 3) in with 1) and 2), since each of the other two grouping possibilities would involve list 4), which has nothing in common with any other list.

Note that grouping on the basis of "words in common" gives us a grouping that we have already decided (above) was unreasonable on intuitive grounds, namely, to group flowerpot with plum and walnut. These sample word lists were fabricated deliberately, not only to illustrate the basic principle by which the lists are grouped, but also to illustrate the apparent weaknesses of the method. We enumerate and discuss these apparent weaknesses in terms of the above sample lists:

*With some assistance from the Encyclopedia Britannica.

A. Word Choices Can Accidentally Relate Documents on Dissimilar Topics. Let us suppose that word list 2) had the word "flower" rather than "blossom," and that (with somewhat greater emphasis on the production of prunes) the word "dry" appeared on the list. We would now have the situation in which lists 2) and 3) would have four words in common, leading to the most unlikely initial grouping of all--plum and flowerpot. Can we permit such subtle shifts in vocabulary and emphasis to have such drastic effects on the outcome of the classification? As we shall eventually see, such inappropriate groupings become less and less likely as 1) the size of the document collection increases, 2) the topical spectrum narrows, and 3) the amount of information (about each document) that is used in grouping is enlarged--i.e., list length is increased.

B. Ties in Number of Words in Common Lead to Instability. Let us assume that lists 2) and 3) had three words in common. Now there is a tie between lists 1) and 3) in their similarity to list 2). In such a case, which group would be formed first, 2) and 3), or 1) and 2)? Since a computer program, unless suitable provision were made, would have no way to decide this issue except through comparison of similarity as we have defined it, a typical program would simply choose the first pair inspected. In other words, we can affect the program's classification simply by physically rearranging the order in which the lists are input. Such instabilities have actually been observed in the computer runs to be described in this paper, but it is not at all clear that this instability is related in any way to the quality or usefulness of the output. We are perhaps uncomfortable with the thought that such instability could lead to many alternative classifications, and that somehow there ought to be only one organization inherent in the document collection. It remains to be seen whether such a viewpoint is really necessary.

C. Raw Lists of Words Omit Semantic Information Which Ought to Affect the Classification. Two important kinds of information omitted would be homography-resolving information and relationship indicators (showing which words on a list are related to each other, and how). An example of both imagined deficiencies is found in the word "plant." In list 2) the word in relation to plums actually refers to the verb "to plant." In list 3) the word is a noun, describing what the flowerpot is to contain, although as far as the information given on the list is concerned, it could be referring to a "plant that manufactures pottery." It could even have both usages in the text of the parent document. The answers to these arguments (tentative answers, admittedly) are that statistical separation of homographs has been shown to occur (8,12), and that relationship indicators--however useful they might be to a user consulting a classification scheme--do not contribute enough information to affect the outcome of the classification significantly. From an information-theory viewpoint, the bulk of the informational bits are contributed by the choices of the words themselves.

Automatic Assignment of Labels to Groups

Four sample word lists have been used in showing the most elementary principles of the Ward grouping procedure, as well as its most apparent possible deficiencies in grouping of word lists. Given that appropriate groups can be formed by such a program, what more can be done? One question is: if we can derive a classification through such statistical procedures, can we also derive labels for the various groups? The answer is that we can, and the mechanism is shown in Figure 2. Six objects are pictured along with their six corresponding attribute lists. The purpose of the diagram is to illustrate that words can be drawn automatically from the attribute lists to give adequate descriptions of the groups, i.e., to describe which common attributes have been most influential in leading to the formation of each group.

The groups of Figure 2 were derived via the same considerations of list similarity that we have already used. The first program pass groups "nail" and "screw," whose lists have five common attributes. On subsequent passes we must lower threshold F to permit the formation of groups of more and more dissimilarity and heterogeneity. On the second pass, "safety pin" and "belt buckle," having four attributes in common, are combined.

On the third pass different things happen, depending on whether one uses the maximum-F or the maximum-A mode of Ward's program. Since I've chosen to use the maximum-F mode, I'll discuss the third pass in those terms. "Poker chip" and "circular cam," having only 2 attributes in common, are paired, whereas the formation of the group of four consisting of "nail," "screw," "safety pin," and "belt buckle," with an average of 3.5 attributes in common, is delayed till the fourth pass; the penalty for reduction of homogeneity, which formation of the group entails, outweighs its lead in average similarity, as may be seen by calculating and comparing values of F for the possible groupings on the third pass. The fifth pass has only one choice, formation of the final group of six.

After the groups are formed, by what rules can we assign labels? Ideally, for any group we would like to select a label that described all and only the members of that group. Our first-formed pair, "nail" and "screw," have the attributes "long" and "headed," which apply to them alone. Each of the other groups of two have at least one such attribute. (In deciding how to specify attributes, I artibrarily distinguished between "cylindrical" and "circular" so that the former could be used to pertain to cross section of structural members and the latter to pertain to gross form.) The group of four has two attributes, "pointed" and "cylindrical," present on all four lists, but not present elsewhere.

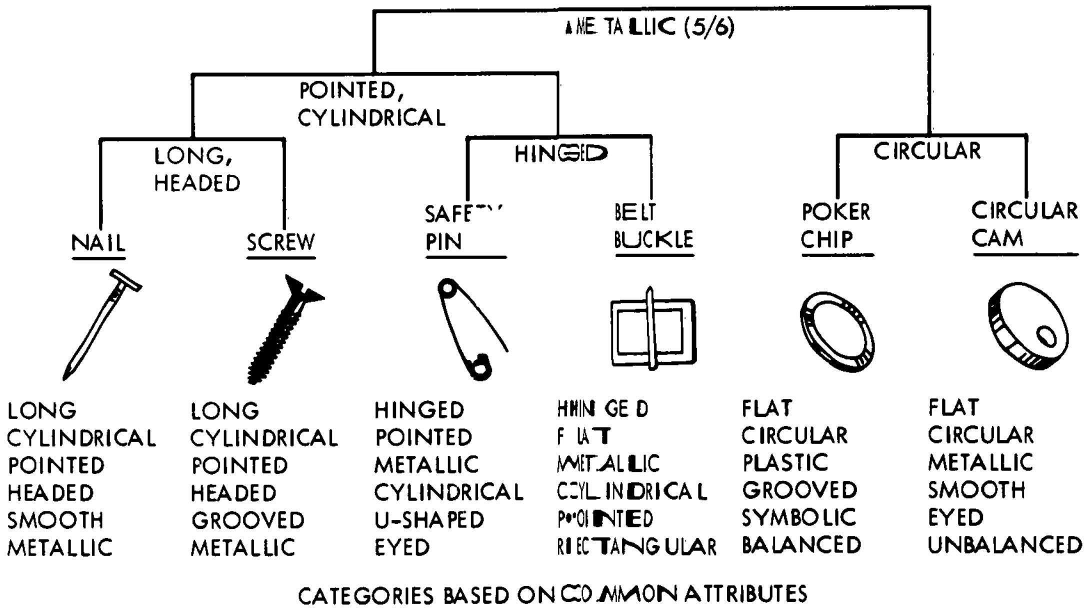


Figure 2. Derivation of Category Labels.

As we ascend upward in the hierarchy, we find some tendency for the attributes to be used up as labels for the smallest categories. There is no attribute, therefore, that perfectly describes the group as a whole. The closest we can come to perfection is "metallic," which describes five out of six of the objects. If the number of objects is increased to the point that five or six levels are generated in the hierarchy, we must either increase the number of attributes per object or else accept group descriptors that do not apply to every group member, or that apply to objects not part of the group. Figure 3 shows a close-up view of the grouping pattern involving seven out of the 100 12-word lists on German affairs, and even though each of the corresponding reports might be said to have "twelve attributes," there are still not many satisfactory choices of labels. The only "perfect descriptor" in Figure 3 is the word "toll," which describes the three members of that group and no outside member.

The notation alongside each label specifies to what extent, if any, the label is not a perfect descriptor of the group. Thus, "allied" describes only 5 out of 8 of the lists in that group (one member of which is not shown), and also describes an additional list at some remote location in the hierarchy; the total number of "allied" tokens is outside of the parentheses, and the fraction of lists described by "allied" is within the parentheses.

As can be seen in Figure 4, which shows the hierarchy* for all 100 reports, there are in general four or five levels; this accounts for part of the difficulty. But also, there is less similarity--on the average--between these lists, even with 12 attributes, than there is between lists in Figure 2.

From a pragmatic viewpoint, a reasonable degree of imperfection of description may not be a serious deficiency. As is well understood in the document retrieval field, there are explicit index tags for a document and there are implicit tags--tags that might well have been chosen to describe the document but were not. Implicit descriptors, unfortunately, are one reason why relevant documents are missed in a search, and this is why people are so interested these days in associative indexing. Thus, though the word "allied" pertains to only five out of eight documents in its group, one can sense that for the documents not tagged by "allied," which--as is seen from Figure 3--are about the various tensions involving East Germany and Berlin, it is reasonable to regard "allied" as one of the implicit tags for those documents. That we should retrieve documents relevant to the term "allied," but not actually bearing the term as a tag, is the whole point of associative indexing. We must take care, of course, not to stretch the "implicit tag" viewpoint too far.

*The smallest categories generally contain two or three--seldom more than four--lists.

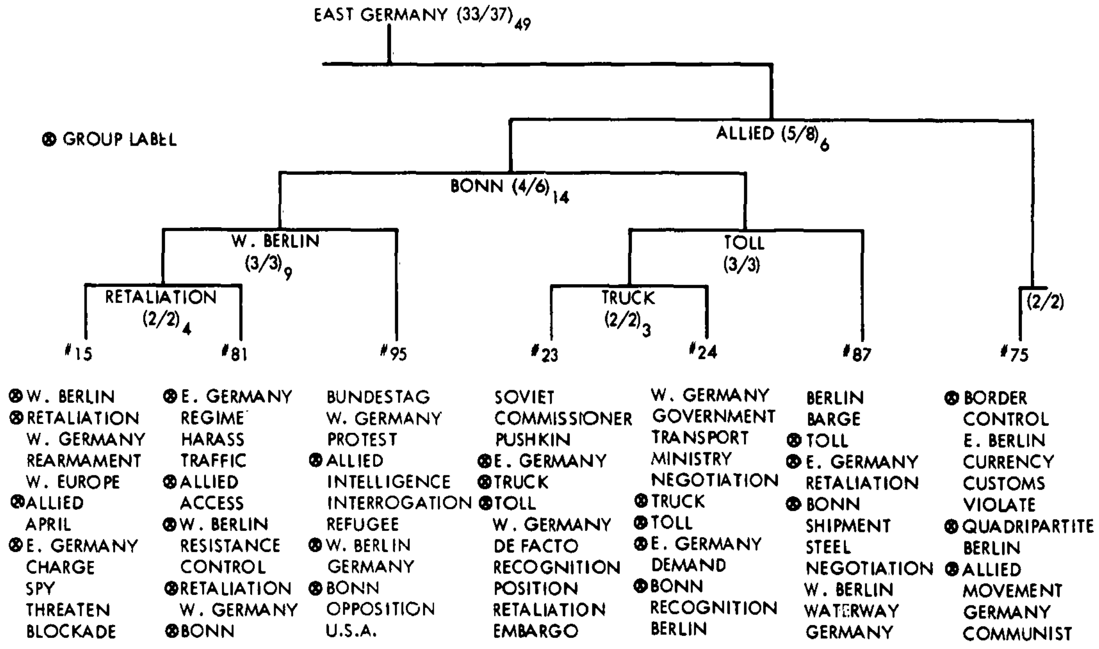


Figure 3. Extent to Which Lists Contain Group Labels.

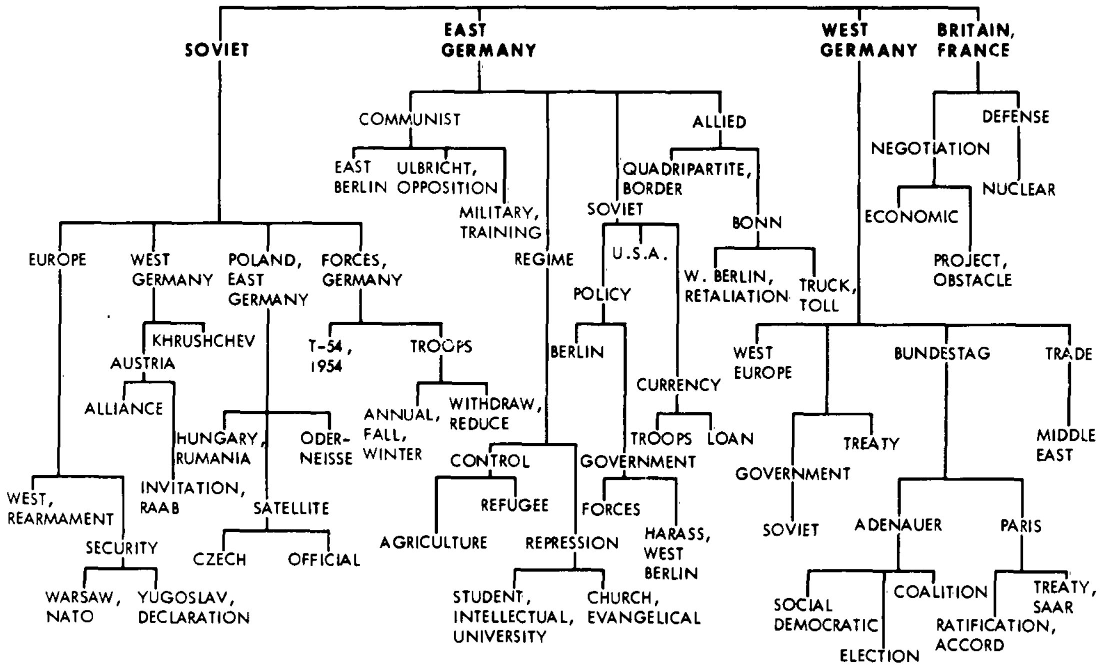


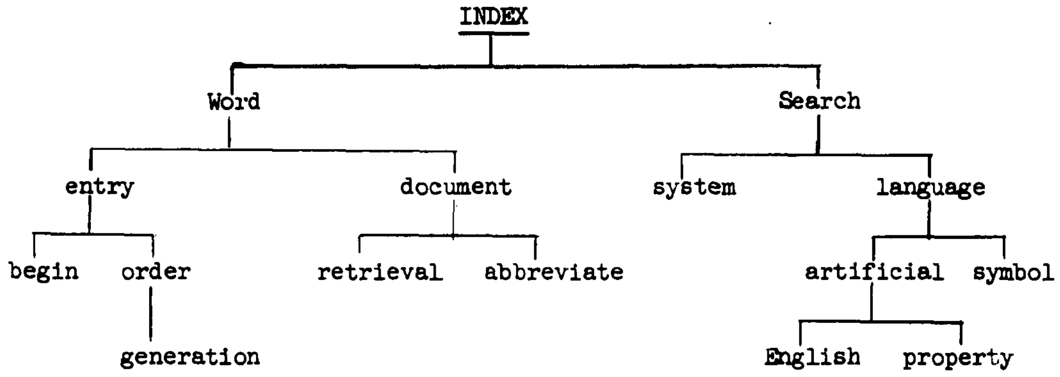
Figure 4. Classification Scheme for 100 Articles Based on Application of Ward's Grouping Program.

The other kind of labelling imperfection--that a given tag describes members outside of the group as well as in it--is even less serious, and in fact may be regarded as not an imperfection at all under conditions of adequate system design. In Figure 4 some words, such as "Soviet," describe several categories and subcategories in different parts of the hierarchy; an alphabetical index of the hierarchy's labels can permit a thorough search of groups described by "Soviet," if such is desired, and could even reference individual documents.

It is in this multiple usage of the same word as a label that we find the homograph-separation power of the Ward grouping procedure. In the third of the three computer runs enumerated earlier, 50 lists in the field of physics and 50 in the field of German affairs were pooled as input to the program. In each field there was substantial usage of the words "satellite" and "force," which are homographs in the true sense of the word. For "satellite" all of the German affairs items used the word to mean "vassal state of the U.S.S.R." All of the physics items used it to mean "man-made earth-circling object." The Ward program not only yielded a perfect separation of reports containing the variant meanings of both "satellite" and "force," but also began the 99th pass with two groups of 50 each--pure physics and pure non-physics.

When one peruses the similarity matrix for all of the lists, however, the clean-cut separation of the two subjects hardly seems miraculous. That half of the matrix describing similarities between individual physics documents and individual German affairs documents contains mostly zeroes. There is a small percentage of document pairs having a similarity of one. When these are looked up they turn out to be tagged by either "force" or "satellite." So there is nothing mysterious about statistical separation of homographs. The reports containing the word "force," in the physical sense, naturally have words in common like "nucleus," "electron," "magnetic," "field," and "charge," and are therefore naturally grouped together by the Ward procedure.

The results of the second run--on 100 lists corresponding to documents in information retrieval--were not so satisfactory as the results for the German reports or for the mixed library just described, chiefly because no words adequately described the largest categories (as in the case of the four major categories of Figure 4). This result is expectable whenever the subject matter in a document collection is too diverse. Another reason for dissatisfaction is vocabulary. A typical structure from the IR hierarchy is:



Alongside of hierarchies containing such crisp words as "Bundestag," "troops," "Khrushchev," "Hungary," and "rearmament," structures such as the above would not seem to shed much light on the organization of the literature in the information retrieval field. I have often contended that the greatest difficulty in retrieving information would be found in information retrieval's own documentation. Nevertheless, even in an area as semantically fuzzy as IR there is great reason for optimism if statistically processed material is touched up with an appropriate amount of post-editing (13).

Earlier in this paper we listed five weaknesses of pure word-grouping and pure document-grouping. It may be evident after the preceding discussion that the Ward grouping procedure is one approach which, with further development, offers great promise of overcoming these weaknesses. It permits:

1) Terse and reasonably accurate labelling of groups of all sizes.

2) Intricate and meaningful organization of groups in relation to each other.

3) Optimum positioning of references to individual documents in a network of descriptive words.

4) Homograph separation and aspect coordination as natural outcomes of the grouping and labelling procedures.

5) A scheme or map that is more easily comprehensible as a result of being analogous to something which is--or could be--a physical arrangement of objects.

REFERENCES

1. Bar-Hillel, Y. "Some Theoretical Aspects of the Mechanization of Literature Searching." Technical Report No. 3., U.S. Office of Naval Research, Washington, D. C., April 1960.

2. Luhn, H. P. "A Statistical Approach to Mechanized Encoding and Searching of Literary Information." IBM Journal, pp. 309-317, October 1957.

3. Doyle, L. B. "The Microstatistics of Text." Information Storage and Retrieval, Vol. 1, pp. 189-214, November 1963.

4. Doyle, L. B. "Indexing and Abstracting by Association." American Documentation, Vol. 13, pp. 378-390, October 1962.

5. Stiles, H. E. "The Association Factor in Information Retrieval." Journal of the ACM, Vol. 8, pp. 271-279, April 1961.

6. Doyle, L. B. "Semantic Road Maps for Literature Searchers." Journal of the ACM, Vol. 8, pp. 553-578, October 1961.

7. Maron, M. E. and Kuhns, J. L. "On Relevance, Probabilistic Indexing, and Information Retrieval." Journal of the ACM, Vol. 8, pp. 216-244, July 1960.

8. Stiles, H. E. "Progress in the Use of the Association Factor in Information Retrieval." Unpublished memorandum, 15 November 1962.

9. Ward, J. H., Jr. and Hook, M. E. "Application of a Hierarchical Grouping Procedure to a Problem of Grouping Profiles." Educational and Psychological Measurement, Vol. 23, pp. 69-82, Spring, 1963.

10. Borko, H. and Bernick, M. D. "Automatic Document Classification: Part II - Additional Experiments." Journal of the ACM, Vol. 11, April 1964, (in Press).

11. Ward, J. H., Jr. "Hierarchical Grouping to Maximize Payoff." WADD-TN-61-29, Wright Air Development Division, Air Research and Development Command, USAF, Lackland Air Force Base, Texas, March 1961.

12. Doyle, L. B. "Statistical Semantics." Information Processing 1962 Proceedings of IFIP Congress, (Munich, 1962), pp. 335-336. Amsterdam: North-Holland Publishing Company, 1963.

13. Doyle, L. B. "Expanding the Editing Function in Language Data Processing." System Development Corporation document, SP-1268, July 1963.

UNCLASSIFIED

System Development Corporation,
Santa Monica, California
SOME COMPROMISES BETWEEN WORD GROUPING
AND DOCUMENT GROUPING.
Scientific rept., SP-1481, by
L. B. Doyle. 26 March 1964, 22p., 4 figs.,
13 refs.

Unclassified report

DESCRIPTORS: Information Retrieval.
Documentation.

Reports that statistical analysis of the text of document collections has yielded, for information retrieval purposes, two broad classes of output: word grouping and document grouping. Also reports that

UNCLASSIFIED

UNCLASSIFIED

associative indexing comes under the general heading of word grouping and automatic classification is a kind of document grouping. States that document grouping and word grouping can be combined to give a scheme of classification with more attractive features than could be achieved with either document grouping or word grouping alone. Discusses a hierarchical grouping program written by J. H. Ward of Lackland Air Force Base for use in classifying personnel by skill and aptitude. States that this program turns out to be nearly ideal as a basis for a mixed document and word-grouping approach.

UNCLASSIFIED